AD-A090 784 HUMAN RESOURCES RESEARCH ORGANIZATION ALEXANDRIA VA F/G 5/10
EXPERIMENTER-CONTROLLED DECISIONS IN THE DESIGN AND ANALYSIS OF--ETC(U)
JUN 65 D REYNOLDS DA-44-188-ARO-2
UNCLASSIFIED NL

1 OF 1
AD A 090784

END
DATE FILMED
11-80
DTIC

LEVEL II

(1)

HumRRO

The George Washington University
HUMAN RESOURCES RESEARCH OFFICE
operating under contract with
THE DEPARTMENT OF THE ARMY

This material has been prepared for review by appropriate research or military agencies, or to record research information on an interim basis.

The contents do not necessarily reflect the official opinion or policy of either the Human Resources Research Office or the Department of the Army.

The Human Resources Research Office is a nongovernmental agency of The George Washington University, operating under contract with the Department of the Army (DA 44-188-ARO-2). HumRRO's mission, outlined in AR 70-8 is to conduct research in the fields of training, motivation, and leadership.

(1)

(9) STAFF PAPER, (12) 59

(6) EXPERIMENTER-CONTROLLED DECISIONS IN THE DESIGN AND ANALYSIS OF PSYCHOLOGICAL RESEARCH.

by

(11) Jun 65

(10) Donald Reynolds

(15) DA-44-1[illegible]-10-[illegible]

This Staff Paper has been prepared for dissemination within HumRRO for purposes of information or coordination internal to the organization. It does not necessarily represent official opinion or policy of either the Human Resources Research Office or the Department of the Army.

DIVISION NO. 1
Human Resources Research Office
The George Washington University
operating under contract with
The Department of the Army

JOB

405260

SP 2
June 1965

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A090784	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* EXPERIMENTER-CONTROLLED DECISIONS IN THE DESIGN AND ANALYSIS OF PSYCHOLOGICAL RESEARCH		5. TYPE OF REPORT & PERIOD COVERED Staff Paper
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Donald Reynolds		8. CONTRACT OR GRANT NUMBER(s) DA 44-188-ARO-2
9. PERFORMING ORGANIZATION NAME AND ADDRESS Human Resources Research Organization 300 North Washington Street Alexandria, VA 22314		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Department of the Army		12. REPORT DATE June 1965
		13. NUMBER OF PAGES 26
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

psychological tests
research management
experimental design
test construction (psychology)
statistical analysis
bias

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This paper reviews certain statistical rationales and procedures that can be useful to research staff in designing and analyzing research. It consists of three parts. The first deals with the necessity of computing the number of subjects required by a given experiment and presents rationales and procedures. The second part explains the advantages of having equal numbers of subjects in experimental treatment 'conditions' or 'cells,' and shows to what extent bias may affect analysis of unequal cell frequencies. The third part outlines the sometimes drastic effect of making multiple comparisons on the same data and suggests alternate procedures.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

Preface

The present paper can best be described as an informal attempt to alert and/or remind research staff members about certain statistical rationales and procedures possibly helpful to them in the design and analysis of research. Since the primary effort involved in this paper is heuristic rather than pedagogical, theoretical notes have been footnoted or referenced rather than dealt with in the body of the paper and an intuitive rather than a derivative approach has guided selection of the expository material.

This paper, like Gaul, is divided into three parts. The first part deals with the necessity of computing the number of subjects required for a given experiment. Rationales and procedures are presented therein. The second part explains the advantages of having equal numbers of subjects in experimental treatment "conditions" or "cells," and shows to what extent bias may enter into the analysis of unequal cell frequencies. The third part outlines the sometimes drastic effect of making multiple comparisons on the same data and suggests some alternate procedures.

It should be pointed out that all three parts deal with decisions solely under the control of the researcher and can be made in advance of data collection. By making decisions of this sort, analysis and interpretation of results can be greatly facilitated.

Comments on the utility of the procedures discussed herein would be appreciated.

Part I - How Many Subjects Do We Need?

Researchers, being human, sometimes forget to have their car serviced, or miss their annual checkup at the physician's or dentist's office. If they were asked to supply the reason(s) for forgetting this "preventive maintenance" on their car, body, or teeth, we would undoubtedly receive quite a range of replies. Such might include some thin rationalizations—easily transparent to the amateur—or might consist of intricately reasoned, balanced structures of thesis, antithesis, and synthesis. All would have in common two contradictory thoughts, however: 1) It is important and, of course, I know about the need for it; and 2) You don't have to make such a fuss over it—it isn't that necessary.

Against this background let us see what "preventive maintenance" as applied to the design of research has to offer. Most researchers will carefully assume the operation of Kelley's Law[1] in the planning of their research. Given this tendency, one finds that large numbers of subjects (Ss) are often run in order for the researcher or experimenter (E) to have the highest confidence in the generality of his findings. But there is Research_1 and Research_2. The former may be considered as the case in which additional data collection is relatively inexpensive—in terms of the E's time and effort, and hidden costs (such as preparation of additional experimental materials, data analysis, overhead, etc.). The latter type of research, Research_2, has the reverse characteristics—extra effort in data collection (such as running "additional" Ss) may be quite costly.

1/ "If anything can go wrong...it will."

In certain cases running many Ss may introduce subtle biases in the data; at the extreme is the case when prolonged observation may result in a change in the phenomenon being studied.

Let us hypothesize a case of $Research_2$. Suppose we have inaugurated a study to determine the effects of "Training Method A" on student proficiency in a course in a particular subject. We choose a USCONARC school as the site of our study, and we take one-half of the input to that school for a given period of time, eventually comparing the effects of the administration of "Training Method A" with the proficiency of the group receiving the conventional training. Quite apart from the design of such a study—which has been dealt with elsewhere (MacCaslin & Cogan, 1964)—what are some of the implications of administering Training Method A to such a large group of Ss? These may be listed as follows:

1. The "experimental" group, by virtue of knowledge of receiving a special or unusual training method, may reach inflated levels of proficiency quite apart from the effects of the method itself. This is the well-known "Hawthorne Effect" (cited in Lindzey, 1954, pp. 1105-1106).

2. The possibility of a reverse effect—performance deficit under "guinea pig" conditions—may not be automatically ruled out, however. This is especially true where data collection covers a prolonged period of time. In such cases the people providing the local

support may lose whatever original enthusiasm they once had, and subtly transmit this loss of enthusiasm to the S-population[2].

3. The cost of running one-half of the student input may be unjustified. Perhaps one-fifth as large a group would have sufficed in order to achieve "statistically significant" results[3]. In the discussion of quality and quantity of support needed that inevitably must occur between HumRRO and USCONARC, the question of "How many subjects do you need?" is a thorny one. It may be just a bit too easy on ourselves to reply, "As many as we can get."

2/ Current events may also be fatal to the prolongation of data collection. An informal report (Ayres, 1964) indicated that correlations of pre- and post-test of supervisory techniques which were averaging in the upper .60's dropped to zero for a group which took its post-test on the afternoon of 22 November 1963, and had listened on the radio during lunch to the events transpiring in Dallas, Texas.

3/ Since the sociology of psychological research is what it is, and since the E is probably going to attempt to publish his findings, he probably will want to indicate the probability of a Type I error, the probability of rejecting the Null Hypothesis when it is in fact true. A second model, based on establishing confidence limits on a difference score (i.e., the difference between a and b is x units or more at the .05 level) also exists but will not be discussed herein.

We are in a much stronger position, operationally speaking, when we can specify the numbers of subjects we need for a given research project. There are several questions one would wish to take into account when thinking about such specification:

1. How large a difference between Group A and Group B would we consider a meaningful difference? This question assumes that the E will statistically test his data at customary levels of significance; the "difference" referred to here is determined on psychological and economic grounds, not on statistical ones.

2. What sort of variance in the measurement of performance (proficiency) can we realistically expect? Here we begin to consider the variability of the two methods to be compared. Will each method yield equally variable data, or is one subject to greater fluctuation? How variable will the yield be?

3. What sort of risk are we willing to take, statistically speaking, in claiming that there is a difference between the groups when there is in fact no difference (and the apparent difference is due to chance or random fluctuation)? This is the risk of a Type I error, the error of rejecting the Null Hypothesis when it is in fact true, as tested by the customary "Level of Significance." It is sometimes also called an "alpha error."

4. What sort of a risk are we willing to take, in analogous manner to Type I error, in claiming that there is no difference between the groups when there is in fact a difference? This is the risk of a Type II error, the error of failing to reject the Null Hypothesis,

when it is in fact false. This error is sometimes called a "beta error." By taking the complement of the Type II error (1.0 minus the Type II error), we obtain the "power" of the statistical test—or the probability that we will not commit a Type II error.

Now what has all of the above to do with research done at HumRRO? In general we have agreed that it would indeed be helpful to know how many subjects we need—if not for the research proper, then for the economics of the research and for the liaison value of this knowledge. One of the purposes of this note is to demonstrate the ease of computing the number of subjects needed, given answers and the above four questions.

In general, an Exploratory Study (ES) precedes the Task Conceptualization Paper (TCP). As set forth above, there are four criteria for determining the number of Ss needed for a given research project. The most difficult question is that posed by Question No. 2, which asks, "What is the expected variance of the observations?" On the basis of the ES, the researcher should have some approximation of the variance; answers to the other questions are set by the E himself following psychological and/or economic rationales. Calculation of the number of Ss required is then a simple and straightforward procedure.

As an illustration, the following example is provided for a two-sided test of significance (where the E cannot or will not predict in advance which group will have a higher "score").

Computing Formula:[4] $$N = \frac{2s^2}{d^2}(z_{\frac{1}{2}a} + z_B)^2$$

Where N = number of subjects needed

1) $2s^2$ = Two times the expected or estimated variance of the observations (proficiency scores, number of items correct on a test of transfer, etc.)

2) d^2 = The smallest difference (squared) between the two groups that the researcher wishes to be able to detect.

3) $z_{\frac{1}{2}a}$ = One-half the usual Level of Significance; if we are working at the p = .05 level, then $z_{\frac{1}{2}a}$, as obtained from tables of percentile values of the normal curve, equals 1.96.

4) z_B = Risk of Type II error; if we wish this to be .10, then z_B = 1.282. This value is obtained analogously from the tables of percentile values of the normal curve.

Note: Both "z-values" ($z_{\frac{1}{2}a}$ and z_B) are always positive.

Example: We have decided, on non-statistical grounds, to look for a difference of 20 units; this is the smallest difference which we would consider as worth finding. We have also decided to use the customary .05 Level of Significance, and wish to limit our chances of making a Type II error to a probability of 10% (or..10). We

---

4/ From Walker & Lev (1953), p. 166. There is an analogous procedure for one-sided tests. The logic of the computation is more fully discussed therein.

have also determined via an ES that the standard deviation will be about 20 units and thus the variance is $(20)^2$ = 400 units (the same units as for the difference above). For a two-sided test which we will eventually make, the values of the above symbols are:

1) = 400

2) = 20^2 = 400

3) = 1.96

4) = 1.282, therefore:

$$N = \frac{2(400)}{400}(1.96 + 1.282)^2 = 2(3.242)^2 = 21.02 \text{ or}$$

N = 22 subjects per group

(Note: This suggests one should choose about 25 Ss per group. The main point here is to determine an order of magnitude—what ballpark do we play in?—rather than absolute numbers.)

Now suppose we had wished to work with different values; let us estimate our variance as 250. Our "smallest significant difference" will now be 10 units. Our Level of Significance will remain the same, but we have decided that we can "afford" to risk a Type II error with a probability of 25% (or .25) this time. What N do we need? We now substitute:

1) 250

2) 10 X 10 = 100

3) 1.96

4) .6745

$$N = \frac{2(250)}{100}(1.96 + .6745)^2$$

$$= 5(2.6345)^2 = 5(6.94) = 34.70 \text{ or}$$

N = 35-40 subjects per group

As a perhaps interesting sidelight, the procedure generalizes to comparisons of more than two groups (the n-group case). The obtained N refers to the number of subjects we need in each group regardless of whether we have two groups and intend to make a t-test, or have n groups and intend to do an Analysis of Variance.[5]

At this juncture, the ease of computation is probably more apparent than the need. Some case histories, a free blend of reality, disguise, and fantasy, are presented below in order to "document" the need.

"Case History #1"

Jim Jones, researcher sans peur, gets into a conversation with Lt. Col. Brown on—of all things—the number of subjects Jones needs at his school. Jones says, hopefully, "As many as we can get, uh, preferably 750 to 1000." Brown: "How long do you plan to use them, and how many at a time?" Jones: "100 per week for 10 weeks, if possible." Brown: "I'm afraid that's out since we can't handle an additional 100 students here without strain. We have to set up support companies, find cooks and bakers, first sergeants, commanding officers, etc. No, (shaking his head) I'm sorry." Jones: "How about 500?" (The rest of the conversation is fairly predictable.)

"Case History #1 Alternate"

Locale and setting as above, with slight modification in past history of events leading up to conversation. Jones: "We've computed the

---

5/ Personal communication, Dr. Eugene Cogan, November 1964.

number of subjects we need, and the minimum number we must have is 440, or 20 subjects in each of 22 groups." Brown: "How long do you need them for?" Jones: "44 per week for 10 weeks, if possible." Brown: "Maybe if we push it, we can arrange to overman a few companies by 20 students for 10 weeks. Are you sure you can't do it with less subjects?" (Jones explains his calculations and goes over the figures with Brown. Brown is reassured and now can use this information in his discussion with Col. Green.)

"Case History #2"

Programed instruction materials are to be prepared to collect data on programing vs. conventional methods of teaching cost accounting. How many booklets should be printed? Jones calculates the number of Ss he needs; adds a percentage for attrition, misprinted booklets, etc.; and presents his figures to his D/R, who carefully notes the thoroughness of Jones. (This is of course reflected in Jones' next year's salary recommendation.)

"Case History #3"

It has been determined that the cost of running a S in Task FIGMENT is $42. Jones wishes to run Ss (using sequential analysis) only until he attains "statistically significant" results; this consideration is based primarily on economic reasons, since adding Ss needlessly rapidly reaches the point of diminished returns. After computation, he determines that the total outlay for data collection is too large and devises alternate ways to collect data, paring costs of collection to $18.75 per S.

"Case History #4"

Jones, having been burnt in the past, decides to run 10 Ss per group in a rather arbitrary way. Remembering Kelley's Law, he computes the N he needs, finding to his horror it is 22 Ss per group. If he ran 10 Ss in each group, he would have virtually no chance of achieving acceptable levels of significance.

Concluding Comments:

For the amount of potential gain, measured against the amount of time taken in the process, one of the best researcher bets is the computation of the number of Ss he needs in his task. Like preventive maintenance, this concept is often widely acknowledged and unwisely avoided. Or forgotten. Like Kelley's Law.

Part II - Why Try for Equality of Cell Frequencies?

It is each researcher's inalienable right to distribute his subjects (Ss) across treatment conditions as he sees fit. But, just as in deviating from the rules of bidding in bridge, he must have some rationale for deviating from the "rules" of assignment of Ss. These "rules," though not particularly stringent, have been established much in the same way as most rules—they have been found to work at an empirical level. The cardinal principle is, of course, to have an equal number of Ss per cell (or treatment). Let us see what happens when this principle is violated.

Let us suppose that we have two groups, one of 15 Ss and the other of 5 Ss. We examine the mean difference between these two groups by means of a t-test, noting that there is heterogeneity of variance. Although the variance from the large group is five times as small as the variance of the small group, we find a "significant" difference at $p = .05$. Surely there is nothing wrong with this! Yet, as examination of Table 1 will show, there is. Our "real" level of significance, as theoretically determined, is $p = .18$ and we have erroneously rejected the Null Hypothesis. Were the train of events to halt there, little would be lost; but we usually take some further action on the basis of our study, making some change in a training program. The ramifications spread. (Note that if we had used 7 Ss in each group—fewer subjects!—we would have been able to reject the Null Hypothesis at the .063 level.)

Now suppose we had two large[6] but unequal groups, one twice the size of the other, and with the variance of the smaller group five times that of the larger group. By examination of Table 2 we can see that our nominal 5% level is in reality a 12% level. And so it goes.

By examination of the two Tables, the reader thus far has discovered for himself three points: (1) A level of significance can be biased on the "safe" side—as when there are two unequal groups and the large variance comes from the large group; (2) A level of significance can be biased on the "unsafe" side—when the larger variance comes from the smaller of two unequal groups; and (3) There need be little or no bias in the level of significance if the sample sizes are equal. Table 3 generalizes the above to the three- and five-sample case.[7]

6/ Most statistical theorists, when they speak of "large n" refer to sample sizes of 20 to 30. Scheffé, however, speaks of "large n" in asymptotic terms; i.e., as n approaches infinity.

7/ Lindquist (1953) speaks of the effects of violating assumptions underlying the F test, citing the Norton study—by now a classic (pp. 78-86). Even when the shapes of the curves and the variances in the n-sample case are very discrepant, normal-theory statistics still provide a remarkably good fit—provided the sample sizes are equal. See also Boneau (1960) in a readable article dealing with these matters applied to t-tests.

Part III - Milking the Data, or
Effects of Making Multiple Tests on the Same Study

In studies of psychological phenomena, we try to arrange things such that we get the best return for our research dollar. This often means that we run an Analysis of Variance (ANOVA) design so that we can assess the effect(s) of the main treatment(s) as well as any interactions that may occur. With ANOVA designs we usually run into the problem of making multiple comparisons—we wish to know which cells of subtreatment combinations are more effective in terms of our independent variable(s). The tendency may arise to "milk the data," i.e., make as many tests of significance as seems suggested by whatever psychological rationales we may have. Even when these decisions are taken a priori, there is an upper limit as to the number of tests we may make. This limit is given by k-1, where k is the number of means (of cells or treatments) obtained in the study. The rationale is best given by Walker and Lev (1953) thusly:

> If as many comparisons are formed as there are degrees of freedom (in this case, k-1), then the sums of squares of a set of orthogonal comparisons constitute a complete subdivision of the total sum of squares. It should be noted that orthogonal sets of comparisons can be made up in an endless number of ways (p. 357; italics and parenthetical comment added).

In other words, the limit on the number of independent comparisons that can be made is given by k-1; even though we may choose to make various

comparisons (perhaps combining certain cell means with those of others prior to comparisons), we are limited in the total number of such comparisons.[8]

Does the foregoing imply that we can promiscuously make t-tests until we reach the magic number of (k-1)? Not at all. For every t-test we make we increase the chances of making a Type I error—of erroneously rejecting the Null Hypothesis. This is intuitively clear when we realize that for every 100 such t-tests we make we shall find significant differences in five of them—if we are working at the 5% level—on the basis of "chance" alone. The probability that one or more of these tests will be "significant" at the 5% level approaches unity rapidly as the number of tests increase. For each test of significance we make, we reduce the level of significance by a factor of $(1\text{-alpha})^h$, where

---

8/ A comment by Dr. Eugene A. Cogan on the above is reproduced in its entirety below. It represents an alternate way of looking at the problem, and the author is grateful to Dr. Cogan for suggesting it.

> "There is, however, another case that is best treated by subdivision of sums of squares. That is, if five groups are run and there is specific interest in comparing Groups One and Two and in comparing Groups Three and Five, and a rather diffuse interest in 'everything else,' it is possible to subdivide the four degrees of freedom into specific tests for the specified effects and also 2 df for 'the rest.' This does not require that significant overall F ratio be evident; in fact, the method suggests we do not even bother to compute the overall F."

h equals the number of comparisons we make and alpha is the level of significance chosen.[9] In order to demonstrate the effects of multiple testing on the same data, Table 4 is offered. This table shows new levels of significance, given a nominal level of 5% and 1%, and represents a straightforward computation of $1-\left[(1\text{-alpha})^{h}\right]$ for h comparisons. It will be noted that this table has the satisfying property of showing that, as h becomes very large, the probability of making a Type I error approaches unity. It also shows that this approach is much less rapid when working at the 1% level.[10]

The problem of multiple comparisons is still not resolved in the psychological literature (Ryan, 1959, 1962; Wilson, 1962). Ryan (1959) lists five different cases in which multiple comparisons are made.

9/ The intuitive rationale underlying the factor of $(1\text{-alpha})^{h}$ has to do with the combination of independent probabilities. "If X_1 and X_2 are independent observations, the joint probability that X_1 will be in C_1 and X_2 in C_2 is the product of their separate probabilities (Walker & Lev, 1953, p. 15)." Since the level of significance—or probability—is unchanged over several comparisons, we multiply the factor by itself for as many comparisons as we make.

10/ The implication seems clear. It is better to avoid fishing expeditions in which trivial comparisons are made along with important ones. By so doing, one (1) has more confidence that Type I error has not been inflated; (2) avoids interpretation of the trivial effects, whether "significant" or not; and (3) focuses attention on the comparisons which are central to the purposes of the study.

Although his paper is restricted to the case in which several different groups are to be compared (as in a simple-randomized design) he has several points which generalize to other kinds of analyses.

He points out that the difference between a priori and a posteriori comparisons is slight. Early workers in the field suggest that when an overall F test is significant one can make t-tests between treatment means; this method would be incorrect if the experimenter had not specified which tests he would make in advance of data inspection. Ryan suggests that this line of reasoning closely parallels that of making one- or two-sided tests:

> ...The one-tailed test is appropriate only if the direction of difference is predicted in advance, and if the experimenter is willing to overlook any difference in the opposite direction, no matter how large. Only two conclusions are possible from the data when a one-tailed test is used—either there is a difference in the predicted direction, or the results of the experiment are inconclusive. In effect, the experiment cannot obtain results which are considered a significant refutation of the prediction. If the experimenter allows for the possibility of a result that contradicts his hypothesis, he must use a two-tailed test, and there is no difference in method of analysis from that....where no predictions are made in advance.

In the case of more than two means, the number of possible conclusions is increased. We may have not only confirmation or

> contradiction of the prediction, but we may also have varying degrees of partial agreement with the prediction... (and) the situation is reduced to essentially the a posteriori case. Only if the experimenter states in advance all possible conclusions and the rules by which these conclusions will be drawn, would he have an a priori test (p. 28; italics added).

The implication is clear: If the experimenter is going to make multiple comparisons after finding a significant F in his Analysis of Variance, he must have specified in advance which t-tests he will make, what conclusions he will draw if only some, all, or none are significant, taking into account all possible ways the results of his tests could come out and the interpretations he would make in each case. (With six means there are five orthogonal comparisons possible, each of which has three possible outcomes: significantly in favor, significantly against, and inconclusive with respect to the hypothesis. This means that the experimenter has to make fifteen interpretations in advance!)

In addition, most workers in the field assume that there is only one type of a Type I error; Ryan points out that there are three, calling these "error rates."

> "1. Error rate per comparison. This is the probability that any one of the comparisons will be incorrectly considered to be significant....
>
> "2. Error rate per experiment. This is the long-run average number of erroneous statements per experiment. In statistical jargon it is the expected number of errors per experiment. Unlike the first error rate, which is a probability, the error rate per experiment could be greater than one. That is, we could set a criterion of "significance" in such a way that we would average three false statements per experiment.

> "3. Error rate experimentwise. This is the probability that one or more erroneous conclusions will be drawn in a given experiment. In other words, experiments are divided into two classes: (a) those in which all conclusions are correct, and (b) those in which some conclusions are incorrect. The error rate experimentwise is the probability that a given experiment belongs in Class (b)." (p. 29; all italics his)

Now we can see why it is incorrect to make k-1 t-tests (on k means) following a significant F at the .05 level. The significant F tells us that our probability of error rate experimentwise, (3) above, is .05, but says nothing about the individual comparisons made. For that error rate we must separately consider the error rate per comparison, (1) above, combining the probabilities of each. Thus, if we make 5 t-tests following a significant F (at the .05 level of significance), the chances are 22.6 out of 100 that we will erroneously reject the Null Hypothesis one or more times (see Ryan, 1959, p. 31), using the experiment as the unit of our analysis.

A Procedural Note

The only safe general procedure for making multiple comparisons would appear to be to use the studentized range test, which is a test referring to a probability distribution of the range of k means, based on an estimated standard error of the mean (Ryan, 1959, p. 43). The studentized range test essentially compares the greatest range in the obtained means with those of the theoretical probability distributions for k means.[11]

11/ Thus keeping the experiment as the unit on which our error rate is based. The studentized range test yields the probability that (at our selected alpha level) one or more comparisons will be significant. (See "Error rate experimentwise," above.)

Tables for the studentized range test may be found in Snedecor (1956, p. 252) and Dixon and Massey (1957, p. 440). In addition, the latter source has a discussion of theory and procedure on pp. 152-155.

Conclusions and Summary

We have briefly reviewed three of the many areas of decision which confront the researcher prior to his assumption of the role of experimenter. He must decide whether or not to compute the number of subjects he will need to attain given levels of significance, how to distribute these subjects across experimental treatments, and finally how many and what kind of statistical analyses to make. Each of these areas will have a different relative weight for different researchers, but all are sources of potential bias and, therefore, must be taken into account.

It was recommended that the number of subjects needed be pre-computed whenever possible. It was also recommended that equal numbers of subjects be assigned to the several treatment conditions (if inequalities across cells exist after data collection due to attrition, random elimination of subjects to reduce to equal cell-frequencies may be possible). It was further recommended that the studentized range test be used whenever multiple comparisons must be made, and certain caveats were noted with respect to multiple testing of experimental hypotheses.

References

AYRES, A.W. Assassination and Assimilation, Amer. Psychologist, 1964, 19, 353.

BONEAU, C.A. The Effects of Violations of Assumptions Underlying the T-Test, Psychol. Bulletin, 1960, 57, 49-64.

DIXON, W.J. and MASSEY, F.J. Introduction to Statistical Analysis. New York: McGraw-Hill, 1957.

EDWARDS, A.E. Experimental Design in Psychological Research. New York: Holt, Rinehart & Winston, 1960 (Rev. Ed.).

EDWARDS, A.E. Statistical Methods for the Behavioral Sciences. New York: Rinehart, 1954.

LINDQUIST, E.F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin, 1953.

LINDZEY, G. Handbook of Social Psychology. Reading, Mass.: Addison-Wesley, 1954.

MACCASLIN, E.F. and COGAN, E.A. Learning Theory and Research Paradigms Applied to Training Research: Some Dissonances, Paper read at American Psychological Convention, Los Angeles, September 6, 1964.

RYAN, T.A. Multiple Comparisons in Psychological Research, Psychol. Bulletin, 1959, 56, 26-47.

RYAN, T.A. The Experiment as the Unit for Computing Ratio of Error, Psychol. Bulletin, 1962, 59, 301-305.

SCHEFFÉ, H. The Analysis of Variance. New York: Wiley and Sons, 1959.

SNEDECOR, G.W. Statistical Methods. Ames, Iowa: Iowa State University Press, 1956.

WALKER, HELEN and LEV, J. Statistical Inference. New York: Holt, Rinehart & Winston, 1953.

WILSON, W. A Note on the Inconsistency Inherent in the Necessity to Perform Multiple Comparisons, Psychol. Bulletin, 1962, 59, 296-300.

WINER, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.

Table 1

New Levels of Significance When Unequal Population Variances Exist Given Small Samples of Unequal Size at a Nominal 5% Significance Level*

Size of Samples	Ratio of Variance 1 to Variance 2						
	1:10	1:5	1:2	1:1	2:1	5:1	10:1
15, 5	23% (Biased on "unsafe" side)	18%	9.8%	5%	2.5% (Biased on "safe" side)	0.8%	0.5%
5, 3	14%	10%	7.2%	5%	3.8%	3.1%	3.0%
7, 6	7%	6.3%	5.8%	5%	5.1%	5.8%	6.3%

*Adapted from Scheffé (1959), p. 353.

Table 2

New Levels of Significance When Unequal Population Variances Exist Given "Large" Sample Sizes at a Nominal 5% Significance Level*

Ratio of Sample 1 to Sample 2	Ratio of Variance 1 to Variance 2				
	1:5	1:2	1:1	2:1	5:1
1:1	5%	5%	5%	5%	5%
2:1	12% (Biased on "unsafe" side)	8%	5%	2.9% (Biased on "safe" side)	1.4%
5:1	22%	12%	5%	1.4%	0.2%

*Adapted from Scheffé (1959), p.340.

Table 3

New Levels of Significance When Unequal Population Variances Exist in Three- and Five-group Cases (n specified, and small) at a Nominal 5% Level of Significance, Using Scheffè's "One-Way Layout" or the Lindquistian "Simple-Randomized" Design

No. of Groups	Ratio of Variances	Group Sizes (n)	New Level of Significance
3	1:2:3	5,5,5	5.6%
		3,9,3	5.6%
		7,5,3	9.2%
		3,5,7	4.0%
3	1:1:3	5,5,5	5.9%
		7,5,3	11.0%
		9,5,1	17.0%
		1,5,9	1.3%
3	25:100:225	3,3,3	7.3% *
		10,10,10	6.6% *+
5	1:1:1:1:3	5,5,5,5,5	7.4%
		9,5,5,5,1	14.0%
		1,5,5,5,9	2.5%

Adapted from Scheffè (1959), p. 354.

* Adapted from Lindquist (1953), p.84.

+ Note how, as n increases, levels of significance tend to revert to the nominal significance level, even with very widely discrepant variances.

Table 4

Values for Actual Level of Significance When Making Multiple Comparisons at a Nominal Level of Significance of 5% and 1% with h =1, 2,....25

Number of Comparisons (h)	For a Nominal 5% Level of Significance	For a Nominal 1% Level of Significance
1	5.0%	1.0%
2	9.8%	2.0%
3	14.3%	3.0%
4	18.6%	3.9%
5	22.6%	5.0%
6	26.5%	5.8%
7	30.2%	6.8%
8	33.7%	7.7%
9	37.0%	8.6%
10	40.1%	9.6%
11	43.1%	10.5%
12	46.0%	11.4%
13	48.7%	12.2%
14	51.2%	13.1%
15	53.7%	14.0%
16	56.0%	14.9%
17	58.0%	15.7%
18	60.3%	16.6%
19	62.3%	17.4%
20	64.2%	18.2%
21	66.0%	19.0%
22	67.6%	19.8%
23	69.3%	20.6%
24	70.8%	21.4%
25	72.3%	22.2%

Note: Computed from the formula $1-\left[(1\text{-alpha})^{h}\right]$. A word on the interpretation of this table is in order. If one makes, for example, 12 comparisons at the 5% level of significance, the risk of his making a Type I error—of falsely rejecting the Null Hypothesis—is 46.0%. Thus, if multiple comparisons must be made, working at the 1% level will insure a margin of safety—up to 5 comparisons.